Another Look at the Data Sparsity Problem

Ben Allison, David Guthrie, and Louise Guthrie

University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
`b.allison@dcs.shef.ac.uk`

Abstract. Performance on a statistical language processing task relies upon accurate information being found in a corpus. However, it is known (and this paper will confirm) that many perfectly valid word sequences do not appear in training corpora. The percentage of n-grams in a test document which are seen in a training corpus is defined as n-gram coverage, and work in the speech processing community [7] has shown that there is a correlation between n-gram coverage and word error rate (WER) on a speech recognition task. Other work (e.g. [1]) has shown that increasing training data consistently improves performance of a language processing task. This paper extends that work by examining n-gram coverage for far larger corpora, considering a range of document types which vary in their similarity to the training corpora, and experimenting with a broader range of pruning techniques. The paper shows that large portions of language will not be represented within even very large corpora. It confirms that more data is always better, but how much better is dependent upon a range of factors: the source of that additional data, the source of the test documents, and how the language model is pruned to account for sampling errors and make computation reasonable.

1   Introduction

In natural language processing, data sparsity (also known by terms such as data sparseness, data paucity, etc) is the term used to describe the phenomenon of not observing enough data in a corpus to model language accurately. True observations about the distribution and pattern of language cannot be made because there is not enough data to see the true distribution. Many have found the phrase "language is a system of very rare events" a notion both comforting and depressing, but fewer have ever seen it as a challenge.

This paper explores the extent to which data sparsity is an issue across a range of test documents and training corpora which vary in size and type. It examines the extent to which the chosen training corpus affects the performance of a task, and how much methods to combat data sparsity (principally smoothing) are responsible for performance. That is, the lower the percentage of a model's parameters which is observed in real language use (the corpus), the higher the percentage of times those parameters must be estimated.

The goal of many language processing tasks is to gather contexts, and build a model of those contexts (e.g. statistical machine translation or automatic speech recognition). To do this, the approach in statistical NLP is typically to gather the necessary information from a corpus, and use the data observed in this corpus to assign a probability distribution. However, as this paper will show, using a very large corpus (1.5 billion words) there are many, many instances in normal language where no probability (or only a zero probability) would be

assigned to a word sequence using the simple distribution derived from the corpus. The majority of these word sequences are legal, but there is insufficient data to estimate their probability.

To combat this problem, smoothing techniques have been proposed. The intuition behind the most popular of these techniques is to take some probability "mass" away from the sequences that have been seen before, so that some can be assigned to the sequences which have not been seen. More complex smoothing approaches interpolate probabilities for an n word sequence with those for its component (*n-k*) word sequences; as explained by [2]:

> "if an *n*-gram has a nonzero count then we use the distribution $\alpha(w_i|w_{i-n+1}^{i-1})$. Otherwise, we backoff to the lower-order distribution $\gamma(w_i|w_{i-n+1}^{i-1})P_{smooth}(w_i|w_{i-n+2}^{i-1})$, where the scaling factor $\gamma(w_i|w_{i-n+1}^{i-1})$ is chosen to make the conditional probabilities sum to one."

This paper quantifies the number of sequences unseen in various documents over a range of corpus sizes (and types), using test documents drawn from a broad range of document types. The question answered is: what percentage of tokens in a new document is unseen in a training corpus? This gives us the number of times that a language model trained in a given domain would have to estimate the probability of an n word sequence without ever having seen that sequence. Clearly, the best language models would minimise the number of times this were necessary. [7] addresses a similar question, considering smaller training corpora and with test documents fixed to those heldout from the training corpus. He investigates the vocabulary size which yields the best word-error rate on a speech recognition task, and shows a correlation between trigram coverage and lower word-error rate.

This paper principally concerns itself with trigram modelling — this is by far the most common n-gram used in modelling, since it is typically considered to provide a good balance between some context and enough repetition to make that context useful. Going beyond this value of *n*, [3] shows that higher-order models (four-grams and above) suffer so much from data sparseness that they become unusable. In the case of trigrams, the probability of any three word sequence W is estimated by:

$$p(W) = p(w_i|w_{i-2}w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})}$$

This paper also explores the number of bigrams and unigrams from new documents which are found in training corpora of up to 1.5 billion words.

The paper also briefly explores some interesting side effects of measuring the percentage of unseen *n*-grams in documents of various types. It shows that different types of documents can be separated by the percentage of their tokens which appear in a fixed corpus, and also shows that domain specific corpora are not always necessary - it depends upon the task in question.

One final area the paper considers is the effect that some common techniques for compression of language models has upon the number of unseen trigram sequences. The results using these techniques indicate that huge improvements in storage requirements for models can be achieved for what some might consider an acceptable loss in observed trigram patterns.

It is almost universally accepted that more data is always better; a phrase in the literature is "There's no data like more data" [5], and [4] suggest "having more training data is normally more useful than any concerns of balance, and one should simply use all the text that is available." [1] showed that performance for their chosen task increased as they increased the size of their training corpus, and furthermore showed that a particular method's relative performance on a small training set will not necessarily be replicated if the training corpus grows in size. However, this paper does not seek to question or necessarily affirm these positions; it is concerned with performance not so much on a specific task, but rather with a problem affecting all tasks using corpora to estimate the probability of word sequences. It seeks to show how often one must make up for deficiencies in various corpora by estimating probabilities of unseen events. The method of these estimations, and their relative performances, is examined in other work; here we are concerned with how often they are necessary.

2   Data Used in the Study

2.1   Training Data

Several corpora were used for training:

The BNC – The British National Corpus is a corpus of over 100 million words of modern English, both spoken and written. It is designed to be balanced by domain and medium (where it was intended to be published. It can be considered to represent most common varieties of modern English language.

The Gigaword Corpus – A large archive of newswire text data acquired by the Linguistic Data Consortium. The total corpus consists of over 1.7 billion words from four distinct international sources of English newswire ranging from approximately 1994-2002.

Medline – 1.2 billion words of abstracts from the PubMed Medline project. Medline was compiled by the U.S. National Library of Medicine (NLM) and contains publications in the fields of life sciences and biomedicine. It contains nearly eleven million records from over 7,300 different publications, spanning 1965 – the present day.

2.2   Testing Data

Initial testing data came from three sources, intended to represent a range of language use. For each source, a collection of documents totalling approximately 6,000 words for each source was produced. Sources were as follows:

Newspaper articles – Documents composed of current stories from
    `http://www.guardian.co.uk` and `http://www.thesun.co.uk`
Scientific writing – From Einstein's Special and General Theory of Relativity
Children's writing – From `http://www.childrens-express.org`, a project producing news and current affairs stories by children, for children

Further testing was performed with large sets of data both to clarify results and to observe the phenomenon that data from "strange" sources had appreciably different patterns of trigram coverage.

Newspaper archives – Documents from the Financial Times archive

Anarchist's Cookbook – Documents from the notorious "Anarchist's Cookbook", originally written during the 1960s (although since updated) and comprising articles for small-scale terrorist acts such as drug production, home-explosive creation and identity fraud [6].

MT system output – Google's attempt to translate Chinese news stories. Includes some manual correction of untranslatable (by Google) characters.

Emails – Messages drawn at random from the Enron email corpus [8].

ASR data – Documents consisting of text output from an ASR system.

3   Method

For the purposes of this paper, coverage is defined as the percentage of tokens from an unseen document found at least once in a training corpus. Both type and token percentages were explored, and token coverage is reported here. For this application, the type/token distinction is as follows: in counting tokens, all instances of a specific n-gram will be counted separately towards the final percentage. For types, each unique n-gram will only be counted once.

Training and test corpora were all prepared in the same way – all non-alphabetic characters were removed, and all words were converted to lower case. For bigrams and trigrams, tokens were formed both by allowing and prohibiting the crossing of sentence boundaries. However, it was found that this had a minimal impact on percentage scores, and the results reported here are those allowing sentence-boundary crossing.

4   Results

Figure 1 shows the average token coverage for the initial test documents of unigrams, bigrams and trigrams against the following corpora:

Corpus 1: 150,000 words from the BNC
Corpus 2: 1million words from the BNC
Corpus 3: 26 million words from the BNC
Corpus 4: Whole (100 million words) BNC
Corpus 5: Gigaword (1.5 billion words)

Figure 2 shows how the coverage of more normal documents degrades significantly when the test documents are from a more unusual source (Medline).

Figure 3 shows the way that different document types separate in terms of their coverage statistics. For each type, between 50 and 250 documents were tested. The figure shows the distribution of coverage scores for these different sources. The types of test document used are indicated in the figure, and for a more complete description, see the "Data" section.

Figure 4 shows the effects of the language model compression techniques on coverage of the same documents as the original tests, using both the BNC and the Gigaword as training corpora. The horizontal axis shows the compression technique, and the vertical, the average coverage (of the same original test documents) using this technique. The techniques used are:
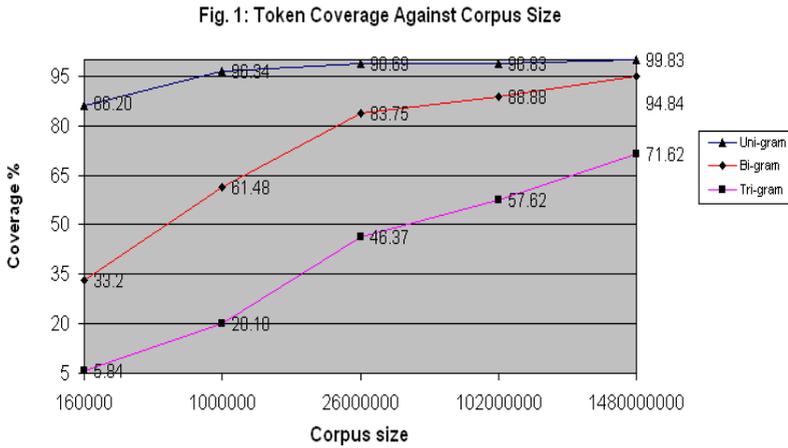
Fig. 1: Token Coverage Against Corpus Size

Fig. 1. Sparsity in initial documents

Fig 2: Coverage of the Gigaword Corpus for Normal and Medline Documents
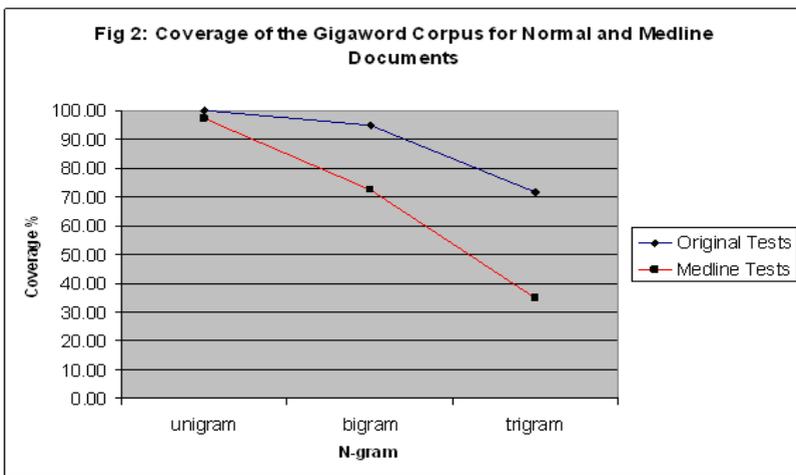
Fig. 2. Gigaword coverage of news and Medline Documents

All trigrams – No compression (as reported above)
Freq 1 – Only trigrams with frequency greater than one are included in the model
Freq 2 – Only trigrams with frequency greater than two are included in the model
Only 50k words – Only trigrams where all three words are one of the 50,000 highest
    frequency words in the corpus are included in the model

   The last two compressions are combinations of the previous ones, e.g. only trigrams with frequency greater than one and all three words are one of the 50,000 highest frequency words in the corpus

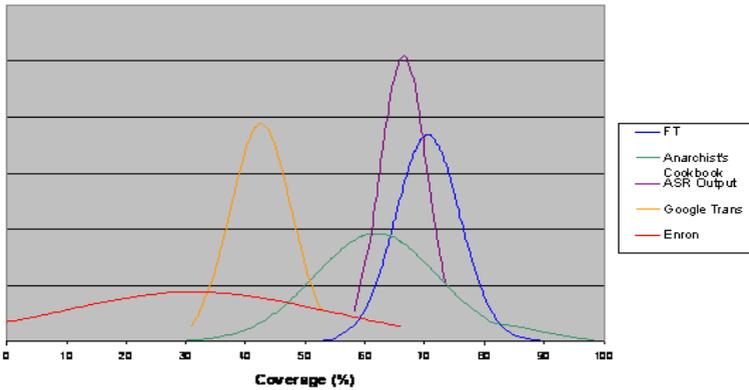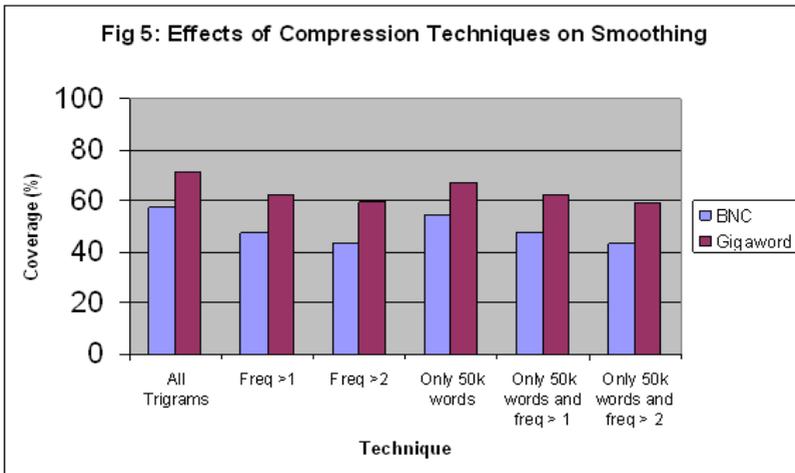**Fig. 3.** Distribution of different document types

Fig. 4. Effects of Compression Techniques

5 Conclusion

The results of these different experiments have shown how varying conditions in a language modelling/context gathering scenario affects the number of unseen events. The higher the frequency of these unseen events, the more reliant any model in these conditions will be upon its method for dealing with unseen events.

The results show a steady coverage increase as the size of the training corpus increases. The largest corpora considered are over one billion words in size, and the results indicate that

one can expect an increase in the coverage up to and beyond this point. Furthermore, they show that in these conditions, with over a billion and a half words of language knowledge at its disposal, a system would still have to estimate the probability of approximately 30% of all legal three word sequences without ever having seen them.

However, some solace can be found in the unigram and bigram coverage rates, which indicate that when a back-off smoothing algorithm must estimate trigram probabilities from bigrams or unigrams, there will be almost no instances where this is not possible due to lack of data. Indeed, 95% of all bigrams can be found in the 1.5 billion word corpus. Where even the bigram is missed, 99.8% of all single words are found within this corpus, meaning the number of times where a word is out-of-vocabulary is tiny. The missed words were either misspellings, colloquialisms which entered language after the corpus creation, or proper nouns. In almost no cases will it be necessary to resort to a last-ditch estimate of a function of the size of the vocabulary.

Various authors have shown a correlation between coverage and language modelling tasks (see [7] for an examination of word-error-rate correlation with coverage in a speech recognition task). However, it is not the purpose of this paper to predict the performance of systems based upon corpora in these conditions, or to evaluate language modelling strategies. Perplexity is not considered here, since it assumes the existence of smoothing approaches in the language model. This paper instead seeks to show the dependency of modelling strategies on their smoothing techniques.

Further results show that the use of domain specific corpora becomes more and more necessary as the n in the n-gram to be modelled increases. Gigaword and Medline do a surprisingly good job of covering one another's unigrams, and arguably do an acceptable job with bigrams. The real unsuitability is evident only when considering trigrams, and one can assume that this phenomenon will grow when considering 4- and 5-grams.

More results show that different document types display different patterns of coverage with respect to a static corpus in the general (large collections), as well as the specific case. These results once again reinforce the hypothesis that, as the domain becomes more unusual with respect to the language model, so the model must more often estimate the probabilities of unseen events. Two of the sources (ASR system output and the Anarchist's Cookbook) are reasonably well approximated by the Gigaword model — both represent well-formed English, if a little unusual, and the ASR system's output is by definition regulated by a language model. The other two sources are less well dealt with by the model — Google's translations clearly represent broken English (as anyone inspecting the system's output will doubtless confirm) and the emails represent direct communication, unlikely to be fitted by a model formed from newswire text.

The last set of results somewhat vindicates compression strategies which throw away infrequent n-grams. The loss in coverage could well be defended in the face of the huge space reduction – from 39 million trigram types for the BNC and 260 million for the Gigaword, down to 3.5 million and 50 million respectively for a moderate ten to fifteen percent drop in coverage. This allows the use of accurate frequency counts obtained from large corpora to be combined with the diminutive size of much smaller resources.

This paper has quantified the reliance of language models on their chosen strategies for estimating probabilities for unseen events. It has shown that in some cases, domain specific corpora are essential, whereas in others they are not so necessary. Finally, it has

given quantifiable defence to some techniques for compressing a language model. As data processing capacities increase over time, the paper gives some evidence that fewer and fewer phrases will have to be estimated by smoothing. Hopefully the weakest link in the language modelling chain will become defunct.

References

1. Banko, M., Brill, E., (2001). Mitigating the Paucity of Data Problem. In Proceedings of the Conference on Human Language Technology.
2. Chen, S., Goodman, J., (1998), An empirical study of smoothing techniques for language modeling. Technical report TR-10-98, Harvard University.
3. Jelinek, F. (1991) Up from trigrams!. In Proceedings Eurospeech '91
4. Manning, C., Schütze, H., (1999) Foundations of Statistical Natural Language Processing. MIT Press.
5. Moore, R. (2001) There's No Data Like More Data (But When Will Enough Be Enough?). In Proceedings of IEEE International Workshop on Intelligent Signal Processing.
6. Powell, W. (1970) The Anarchist's Cookbook. Ozark Pr Llc.
7. Rosenfeld, R., (1995), Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data. In Proceedings Eurospeech '95.
8. Klimt, B. Yang, Y. (2004) Introducing the Enron Email Corpus. Carnegie Mellon University.